

INFERENCE OF CANCER PROGRESSION MODELS WITH BIOLOGICAL NOISE: SUPPLEMENTARY MATERIALS

ILYA KORSUNSKY, DANIELE RAMAZZOTTI, GIULIO CARAVAGNA, BUD MISHRA

1. DETAILED COMPARISON OF PERFORMANCE RESULTS ON SYNTHETIC DATA

Here, we include the performance results for the comparison of POLARIS to the optimization BIC and the clairvoyant DiProg. Figures 1, 2, and 3 show the comparison results using recall and precision as performance metrics and both small and asymptotic sample sizes, for CMPNs, DMPNs, and XMPNs, respectively. We separated the recall and precision in order to highlight the asymmetry in POLARIS's performance. That is, POLARIS performs considerably better in recall and consistently introduces a slightly higher number of false edges in the reconstructed graph. The asymptotic sample size is included to experimentally verify the convergence of POLARIS. Note that theorem 1 only guaranteed convergence on graphs without transitive edges, but even with transitive edges, POLARIS converges almost completely at only 2000 samples.

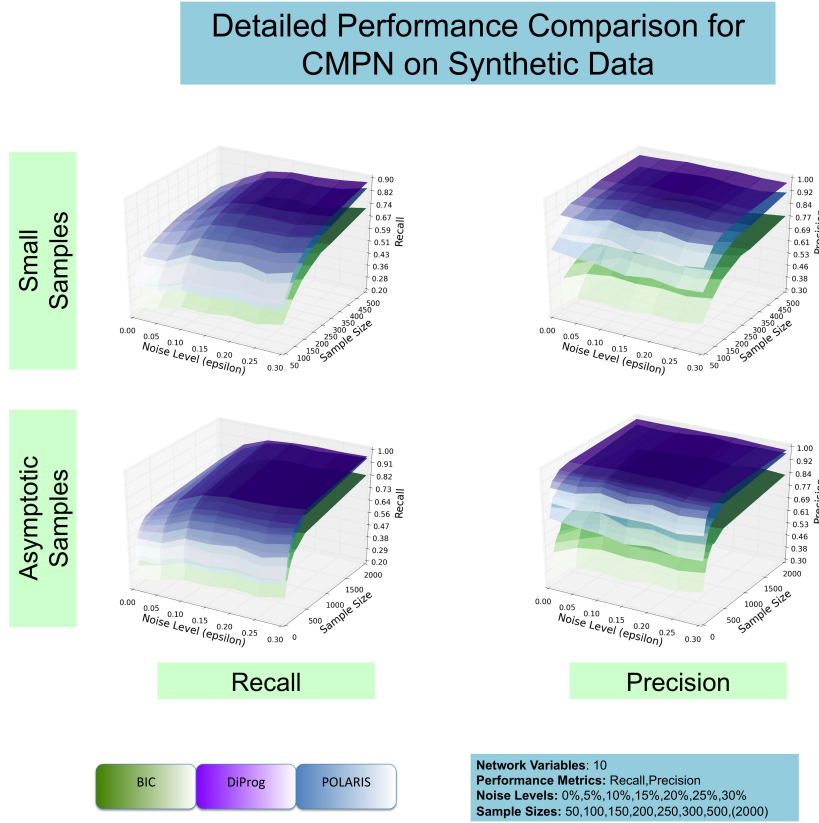


FIGURE 1. The experimental performance results for POLARIS, BIC, and clairvoyant DiProg on CMPNs, measured in terms of recall (*left panels*) and precision (*right panels*). To show the asymptotic behavior of the three algorithms, we plotted the performance for sample sizes up to 2000 (*bottom panels*). For comparison, we also included the performance on more realistic sample sizes (*top panel*).

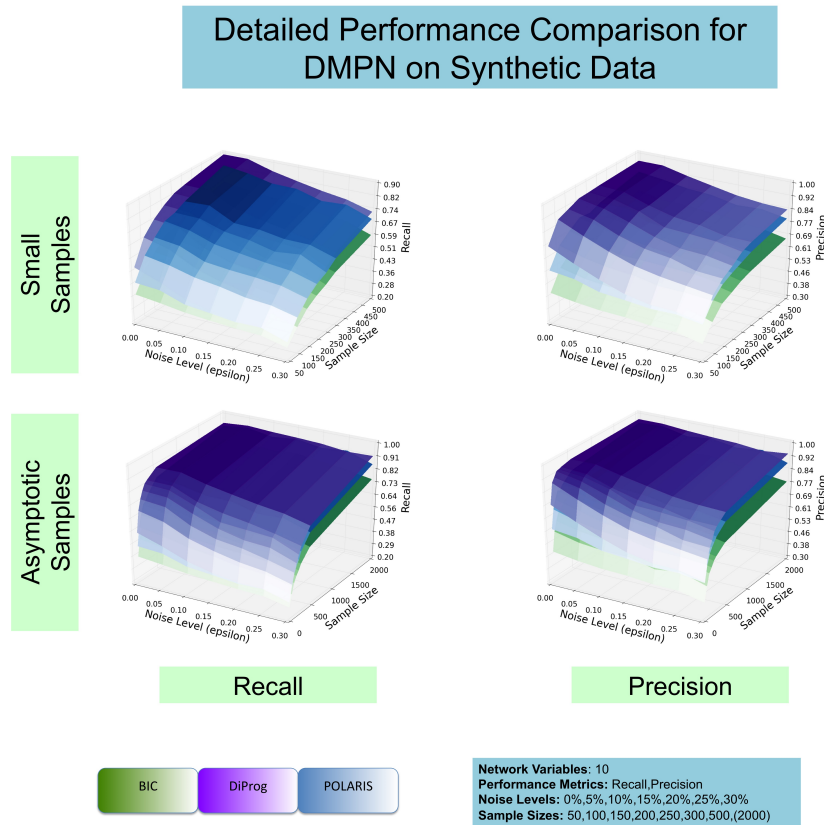


FIGURE 2. The experimental performance results for POLARIS, BIC, and clairvoyant DiProg on DMPNs, measured in terms of recall (*left panels*) and precision (*right panels*). To show the asymptotic behavior of the three algorithms, we plotted the performance for sample sizes up to 2000 (*bottom panels*). For comparison, we also included the performance on more realistic sample sizes (*top panel*).

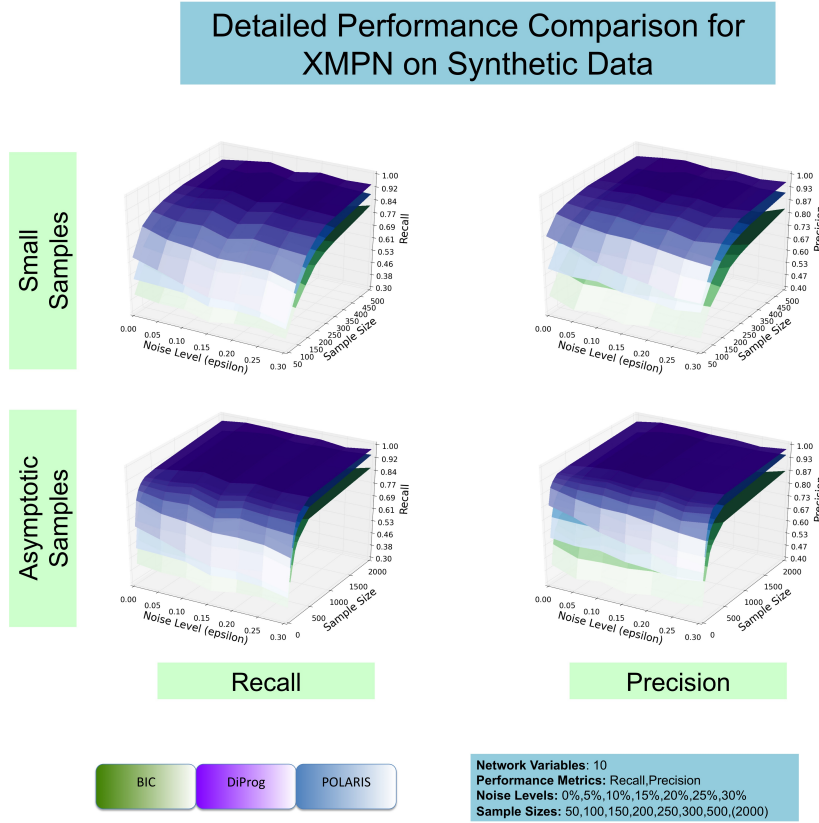


FIGURE 3. The experimental performance results for POLARIS, BIC, and clairvoyant DiProg on XMPNs, measured in terms of recall (*left panels*) and precision (*right panels*). To show the asymptotic behavior of the three algorithms, we plotted the performance for sample sizes up to 2000 (*bottom panels*). For comparison, we also included the performance on more realistic sample sizes (*top panel*).

Figure 4 demonstrates the efficacy and correctness of the α -filter in rejecting hypotheses prior to optimization of the score, in each of the three types of MPNs. For each type of MPN, the average number of rejected true hypotheses is considerably smaller than one and converges to zero for medium sample sizes. The α -filter is particularly effective at pruning the hypothesis space of XMPNs, rejecting approximately 1000 hypotheses on average, out of a possible 1300 hypotheses. It is slightly less effective for CMPNs, rejecting between 500 and 1000 hypotheses. Finally, it is least effective for DMPNs, rejecting between 150 and 350 hypotheses.

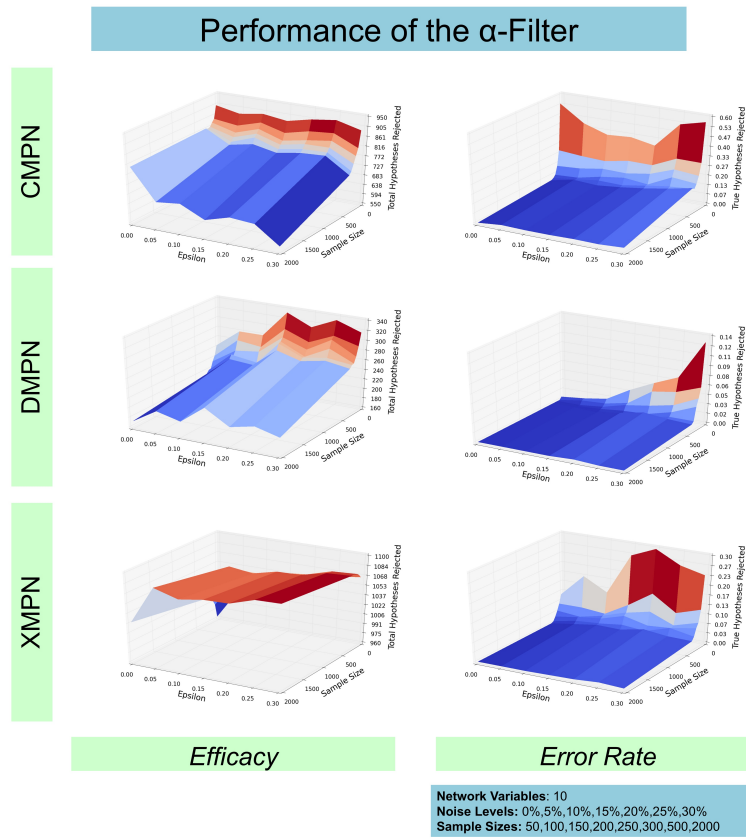


FIGURE 4. The α -filter rejects hypotheses prior to optimization of the score. The figures on the left show the efficacy, measured in terms of the number of hypotheses eliminated prior to optimization. The figures on the right show the error rate, measured in terms of the average number of true hypotheses rejected.

2. TIME COMPLEXITY OF POLARIS OPTIMIZATION

The evaluation of POLARIS scores for all hypotheses dominate the computational complexity of our algorithm. We analyze the asymptotic complexity of this computation and show that its parametric complexity is exponential, where the exponent is determined by the parameter. For a fixed (in practice, small) value of the parameter, POLARIS is polynomial and tractable. To estimate the complexity, we first determine the complexity of computing the score for any single hypothesis; then we multiply this function by the number of hypotheses to get the total cost, which is

$$O(M \cdot N^2 \cdot (N - 1)^k).$$

Here, the parameter k is the maximum number of parents for any node (and can be safely bounded by 3, in practice), and the input size is determined by M and N : respectively, the number of samples, and the number of variables. In practice, the α filter helps performance tremendously, as it avoids the log likelihood (LL) computation for at least nearly half of the hypotheses (see figure 4).

2.0.1. Computing the score for a single hypothesis. The bulk of the score computation effort is expended in computing α and the LL. The α computation is divided into computing θ_i^+ 's and θ_i^- 's, which are just the probabilities of each row in the matrix, encoding Conditional Probability Distributions, CPD. Both computations entail counting the number of samples that correspond to each row and thus in total, take $O(M \cdot N)$ time. The maximum likelihood (ML) parameters in the LL score are precisely the θ_i^+ 's and θ_i^- 's computed for α . Actually computing the LL given the ML parameters requires iterating through the samples one more time and matching each sample to its corresponding CPD row. Thus, LL computation also takes $O(M \cdot N)$ time. Combining all, the total local score computation for one node still takes $O(M \cdot N)$ time.

2.0.2. Number of hypotheses. The hypotheses corresponding to one node consist of its possible parent sets. A node can have parent sets of size 0 to size k , but it cannot be its own parent. Thus, the total number of parent sets for one node is $\sum_{i=0}^k \binom{N-1}{i}$. The final term dominates the series, and thus asymptotically, the number of hypotheses for one node is $O(N^k)$.

3. PROOFS OF THEOREMS ON ASYMPTOTIC CONVERGENCE

Next, in this section, we prove several important properties about the asymptotic performance of POLARIS. The main results are summarized in Theorem 1, which defines the type of structures that are learnable by POLARIS and the conditions under which they are guaranteed to be learnable.

Lemma 1 (Convergence of α -filter). *For a sufficiently large sample size, M , the α -filter produces no false negatives for Conjunctive, Disjunctive and Exclusive Disjunctive Monotonic Progressive Networks: CMPNs, DMPNs, and XMPNs, respectively.*

Proof:

By the law of large numbers, the empirical estimates for all rows of the CPDs will converge to their corresponding true parameter values. To show that the α filter will not create false negatives, we show that α for all true parent sets must be strictly positive for all rows of the CPDs. The α values for positive rows are always 1 and will thus never be negative. The α values for negative rows may be negative, if $\theta^+ < \theta_i^-$, for negative row i of a CPD and θ^+ as appropriately defined for each of the MPN types. Thus, we will show that for all 3 types of MPNs, each negative row will have a strictly positive α . In all three cases, we use the fact that the conditional probability for all negative rows of all CPDs is strictly below ϵ and that for the positive rows is strictly above ϵ .

Case I: CMPN. $\theta^+ = \mathcal{P}(X = 1 \mid \sum Pa(X) = |Pa(X)|)$. Here, θ^+ refers to the conditional probability of 1 positive row, which is by definition larger than ϵ , or restated, $\theta^+ - \epsilon > 0$. Combined with the fact that $\theta^- < \epsilon$, it follows that $\theta^+ > \theta^-$ and thus, α will never be negative.

Case II: DMPN. $\theta^+ = \mathcal{P}(X = 1 \mid \sum Pa(X) > 0)$. The derivation below establishes that θ^+ is always strictly larger than ϵ for the true parents sets in a DMPN. The summation in step (1) is over all values of the parents that are not all zeroes. Here, n refers to the number of parents in $Pa(X)$. That is, $n = |Pa(X)|$. The inequality in step (2) exploits the fact that each conditional probability $\mathcal{P}(X \mid \sum Pa(X) = i)$ corresponds to a positive row and is thus strictly larger than ϵ .

$$\begin{aligned}
& \mathcal{P}(X \mid \sum Pa(X) > 0) \\
&= \frac{\mathcal{P}(X, \sum Pa(X) > 0)}{\mathcal{P}(\sum Pa(X) > 0)} \\
&= \frac{\sum_{i=1}^{2^n-1} \mathcal{P}(X, \sum Pa(X) = i)}{\sum_{i=1}^{2^n-1} \mathcal{P}(\sum Pa(X) = i)} \quad [\text{step(1)}] \\
&= \frac{\sum_{i=1}^{2^n-1} \mathcal{P}(X \mid \sum Pa(X) = i) \mathcal{P}(\sum Pa(X) = i)}{\sum_{i=1}^{2^n-1} \mathcal{P}(\sum Pa(X) = i)} \\
&> \frac{\sum_{i=1}^{2^n-1} \epsilon \mathcal{P}(\sum Pa(X) = i)}{\sum_{i=1}^{2^n-1} \mathcal{P}(\sum Pa(X) = i)} \quad [\text{step(2)}] \\
&= \epsilon \cdot \frac{\sum_{i=1}^{2^n-1} \mathcal{P}(\sum Pa(X) = i)}{\sum_{i=1}^{2^n-1} \mathcal{P}(\sum Pa(X) = i)} = \epsilon
\end{aligned}$$

Case III: XMPN. $\theta^+ = \mathcal{P}(X = 1 \mid \sum Pa(X) = 1)$. The derivation below shows, just like in the DMPN, that $\theta^+ > \epsilon$ for all true parents sets in the XMPN. The reasoning behind the steps is similar to that above, except for the summation in step (2) is over the rows in which exactly one parent takes value 1 and the rest take value 0. To denote this, we use the standard notation $Pa_i(X)$ to mean the i^{th} parent of X and $Pa_{-i}(X)$ to mean all parents except for the i^{th} parent of X .

$$\begin{aligned}
& \mathcal{P}(X \mid \sum Pa(X) = 1) \\
&= \frac{\mathcal{P}(X, \sum Pa(X) = 1)}{\mathcal{P}(\sum Pa(X) = 1)} \\
&= \frac{\sum_{i=1}^n \mathcal{P}(X, Pa_i(X) = 1, Pa_{-i}(X) = 0)}{\sum_{i=1}^n \mathcal{P}(Pa_i(X) = 1, Pa_{-i}(X) = 0)} \quad [\text{step(1')}] \\
&= \frac{\sum_{i=1}^n [\mathcal{P}(X \mid Pa_i(X) = 1, Pa_{-i}(X) = 0) \mathcal{P}(Pa_i(X) = 1, Pa_{-i}(X) = 0)]}{\sum_{i=1}^n \mathcal{P}(Pa_i(X) = 1, Pa_{-i}(X) = 0)} \\
&> \frac{\sum_{i=1}^n \epsilon \mathcal{P}(Pa_i(X) = 1, Pa_{-i}(X) = 0)}{\sum_{i=1}^n \mathcal{P}(Pa_i(X) = 1, Pa_{-i}(X) = 0)} \\
&= \epsilon \cdot \frac{\sum_{i=1}^n \mathcal{P}(Pa_i(X) = 1, Pa_{-i}(X) = 0)}{\sum_{i=1}^n \mathcal{P}(Pa_i(X) = 1, Pa_{-i}(X) = 0)} = \epsilon. \quad \square
\end{aligned}$$

Lemma 2(Consistency of POLARIS). *POLARIS is a statistically consistent score.*

Proof:

Let M be the number of samples generated by the graph $G^* = (V, E^*)$. Let $G = (V, E)$ be the graph learned by maximizing the POLARIS score, and G_{BIC} be the graph learned by maximizing the BIC score, both for a sufficiently large M . The POLARIS score consists of three terms: the log-likelihood (LL) term and the regularization term from BIC and the monotonicity term. Each of these terms grows at different rates. The LL term grows linearly ($O(M)$) with the number of samples. The regularization term grows logarithmically ($O(\log M)$). The monotonicity term does not grow ($O(1)$), since the sum of α scores ($\sum_{d \in D} \alpha_d$) grows linearly with the number of samples, M , but it is weighted by $1/M$. Consequently, it is subsumed by the other two terms. Thus, any perturbation to the graph G that would increase the monotonicity score but decrease the BIC score would also decrease the POLARIS score. From the consistence of BIC theorem, we know that any perturbation to the undirected skeleton or v -structures of G_{BIC} would result in a lower BIC score. It follows that for sufficiently large M , the addition of the monotonicity term will not change the undirected skeleton or v -structures of G_{BIC} . Therefore, G is I -equivalent to G_{BIC} and by transitivity, G is I -equivalent to G^* \square .

Theorem 1 (Convergence conditions for POLARIS). *For a sufficiently large sample size, M , under the assumptions of no transitive edges and faithful temporal priority relations between nodes and their parents at least for nodes that have exactly one parent, optimizing POLARIS converges to the exact structure for MPNs. Proof:*

Let $G^* = (V, E^*)$ be the graph that generates the data and G , the graph learned by optimizing the POLARIS score. By the POLARIS consistency Lemma, for sufficiently large M , the undirected skeleton and v -structures of G are the same as those of G^* . Below, we show that under assumptions of temporal priority for all parent-child relations, $G = G^*$. We proceed by showing that the parent set of each node is learned correctly, by considering

nodes that have zero parents, one parents, or two or more parents. It then follows that all of the edges in the undirected skeleton of G^* are oriented correctly and thus $G = G^*$.

Case I: X_i has 0 parents. If X_i has no parents, then the undirected skeleton around X_i will only include the edges to the children of X_i . Thus, the empty parent set is learned correctly.

Case II: X_i has 1 parent. Let X_j be the parent of X_i .

Case IIA: X_j has 0 parents. By definition, X_j has 0 parents and X_i has exactly 1 parent, X_j . Reorienting the edge $X_j \rightarrow X_i$ to $X_j \leftarrow X_i$ results in an I -equivalent graph globally, because the edge is not involved in a v -structure in either orientation. Thus, the BIC score for both orientations is the same, and in order for POLARIS to correctly choose $X_j \rightarrow X_i$ over $X_i \rightarrow X_j$, it must be the case that $\alpha_{X_i \rightarrow X_j} < \alpha_{X_j \rightarrow X_i}$. In the derivation below, we show that this condition is equivalent to the condition for temporal priority. Namely, $\alpha_{X_i \rightarrow X_j} < \alpha_{X_j \rightarrow X_i}$ is equivalent to $\mathcal{P}(X_i) < \mathcal{P}(X_j)$. To conserve space, we let $\mathcal{P}(X_i | X_j) = \theta^+$ and $\mathcal{P}(X_i | \bar{X}_j) = \theta^-$. Also, we use the identity $\mathcal{P}(X_i) = \mathcal{P}(X_i | X_j)\mathcal{P}(X_j) + \mathcal{P}(X_i | \bar{X}_j)\mathcal{P}(\bar{X}_j) = \theta^+\mathcal{P}(X_j) + \theta^-\mathcal{P}(\bar{X}_j)$. The following statements are all equivalent

$$\begin{aligned}
& \alpha_{X_i \rightarrow X_j} < \alpha_{X_j \rightarrow X_i} \\
& \equiv \frac{\mathcal{P}(X_j | X_i) - \mathcal{P}(X_j | \bar{X}_i)}{\mathcal{P}(X_j | X_i) + \mathcal{P}(X_j | \bar{X}_i)} < \frac{\theta^+ - \theta^-}{\theta^+ + \theta^-} \\
& \equiv \frac{\theta^+ \frac{\mathcal{P}(X_j)}{\mathcal{P}(X_i)} - (1 - \theta^+) \frac{\mathcal{P}(X_j)}{\mathcal{P}(X_i)}}{\theta^+ \frac{\mathcal{P}(X_j)}{\mathcal{P}(X_i)} + (1 - \theta^+) \frac{\mathcal{P}(X_j)}{\mathcal{P}(X_i)}} < \frac{\theta^+ - \theta^-}{\theta^+ + \theta^-} \\
& \equiv \frac{\frac{\theta^+}{\mathcal{P}(X_i)} - \frac{1 - \theta^+}{1 - \mathcal{P}(X_i)}}{\frac{\theta^+}{\mathcal{P}(X_i)} + \frac{1 - \theta^+}{1 - \mathcal{P}(X_i)}} < \frac{\theta^+ - \theta^-}{\theta^+ + \theta^-} \\
& \equiv \frac{\theta^+(1 - \mathcal{P}(X_i)) - (1 - \theta^+)\mathcal{P}(X_i)}{\theta^+(1 - \mathcal{P}(X_i)) + (1 - \theta^+)\mathcal{P}(X_i)} < \frac{\theta^+ - \theta^-}{\theta^+ + \theta^-} \\
& \equiv \frac{\theta^+ - \mathcal{P}(X_i)}{\theta^+ - 2\theta^+\mathcal{P}(X_i) + \mathcal{P}(X_i)} < \frac{\theta^+ - \theta^-}{\theta^+ + \theta^-},
\end{aligned}$$

which is equivalent to the following inequalities:

$$\begin{aligned}
& \frac{\theta^+ - (\theta^- + \mathcal{P}(X_j)(\theta^+ - \theta^-))}{\theta^+ - 2\theta^+(\theta^- + \mathcal{P}(X_j)(\theta^+ - \theta^-)) + \theta^- + \mathcal{P}(X_j)(\theta^+ - \theta^-)} < \frac{\theta^+ - \theta^-}{\theta^+ + \theta^-} \\
& \equiv \frac{(\theta^+ - \theta^-)(1 - \mathcal{P}(X_j))}{\theta^+ - 2\theta^+\theta^- - 2\theta^+\mathcal{P}(X_j)(\theta^+ - \theta^-) + \theta^- + \mathcal{P}(X_j)(\theta^+ - \theta^-)} < \frac{\theta^+ - \theta^-}{\theta^+ + \theta^-} \\
& \equiv \frac{1 - \mathcal{P}(X_j)}{\theta^+ - 2\theta^+\theta^- - 2\theta^+\mathcal{P}(X_j)(\theta^+ - \theta^-) + \theta^- + \mathcal{P}(X_j)(\theta^+ - \theta^-)} < \frac{1}{\theta^+ + \theta^-},
\end{aligned}$$

thus implying

$$\begin{aligned}
& \theta^+ - 2\theta^+\theta^- - 2\theta^+\mathcal{P}(X_j)(\theta^+ - \theta^-) + \theta^- + \mathcal{P}(X_j)(\theta^+ - \theta^-) > (1 - \mathcal{P}(X_j))(\theta^+ + \theta^-) \\
& \equiv \theta^+ - 2\theta^+\theta^- - 2(\theta^+)^2\mathcal{P}(X_j) + 2\theta^+\theta^-\mathcal{P}(X_j) + \theta^- + \theta^+\mathcal{P}(X_j) - \theta^-\mathcal{P}(X_j) \\
& > \theta^+ + \theta^- - \theta^+\mathcal{P}(X_j) - \theta^-\mathcal{P}(X_j).
\end{aligned}$$

Simplifying further, we have

$$\begin{aligned}
& -2\theta^- - 2\theta^+\mathcal{P}(X_j) + 2\theta^-\mathcal{P}(X_j) > -2\mathcal{P}(X_j) \\
& \equiv \theta^- + \theta^+\mathcal{P}(X_j) - \theta^-\mathcal{P}(X_j) < \mathcal{P}(X_j) \\
& \equiv \theta^+\mathcal{P}(X_j) + \theta^-(1 - \mathcal{P}(X_j)) < \mathcal{P}(X_j) \\
& \equiv \mathcal{P}(X_i) < \mathcal{P}(X_j).
\end{aligned}$$

Case IIB: X_j has 1 or more parents. Incorrectly reorienting the edge $X_j \rightarrow X_i$ to $X_j \leftarrow X_i$ makes X_i a parent of X_j . Because G^* is acyclic and has no transitive edges, there are no edges between X_i and the true parents of X_j . Thus, making X_i a new parents of X_j creates a new v -structure (case III proves that if X_j has 2 or more parents, then they are all unwed), consisting of X_i , X_j , and the true parents of X_j , that is not in G^* . This contradicts the consistency of POLARIS and thus the edge $X_j \rightarrow X_i$ will never be reoriented.

Case III: X_i has 2 or more parents. Because G^* has no transitive edges, there cannot be any edge between any two parents of X_i . Thus, the parents of X_i are unwed and form a v -structure with X_i . Because POLARIS is consistent, this v -structure is learned correctly. \square .

Corollary 1 (Convergence conditions for POLARIS with filtering). *For a sufficiently large sample size, M , under the assumptions of no transitive edges and faithful temporal priority relations, filtering with the α -filter and then optimizing POLARIS converges to the exact structure for MPNs. Proof:*

In Lemma 1, we showed that α -filtering removes no true parent sets. In Theorem 1, we showed that given a hypothesis space that includes the true parent sets, optimizing POLARIS returns the true graph. Because the α -filter does not remove the true parent sets from the hypothesis space, optimizing POLARIS will still return the correct structure on the filtered hypothesis space. \square .